An Analysis of Collaborative Patterns in Large-Scale Ontology Development Projects

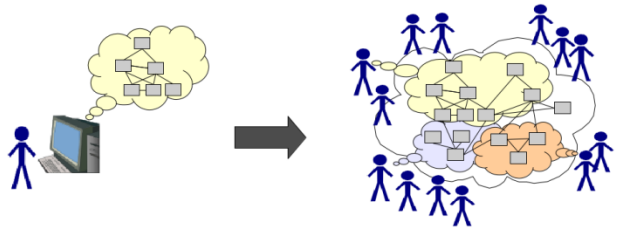
Sean Falconer, Tania Tudorache, Natasha Noy
Stanford Center for Biomedical Informatics Research

K-CAP 2011
Banff, Alberta, Canada

June 27, 2011

Motivation

- Large scale ontologies are developed in groups
- Several ontology editors support collaboration
- Learn about how collaborative ontology development projects work
- What are the **similarities and differences** in terms of **user roles**?
- What characteristics distinguishes the **role of a contributor**?
- Is there a relationship to roles in open source development or Wikipedia? Can methods and tools be borrowed from these communities?
- Can a deeper understanding of these projects and roles **inform tool and ontology development**?

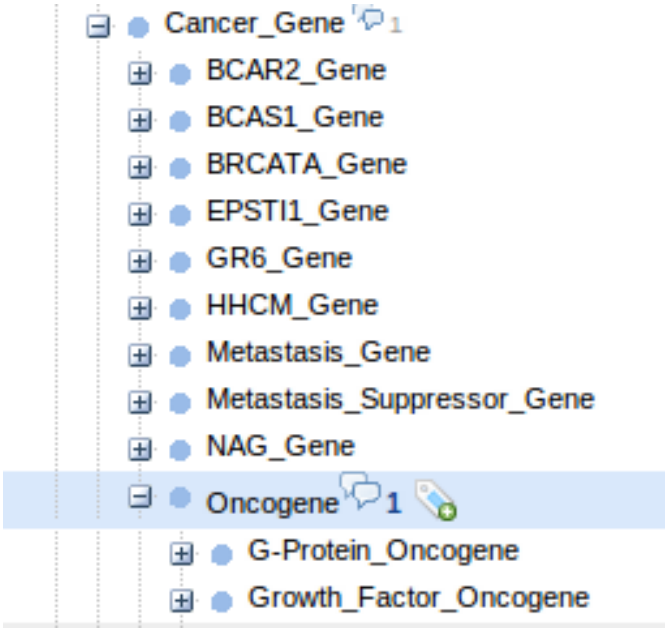


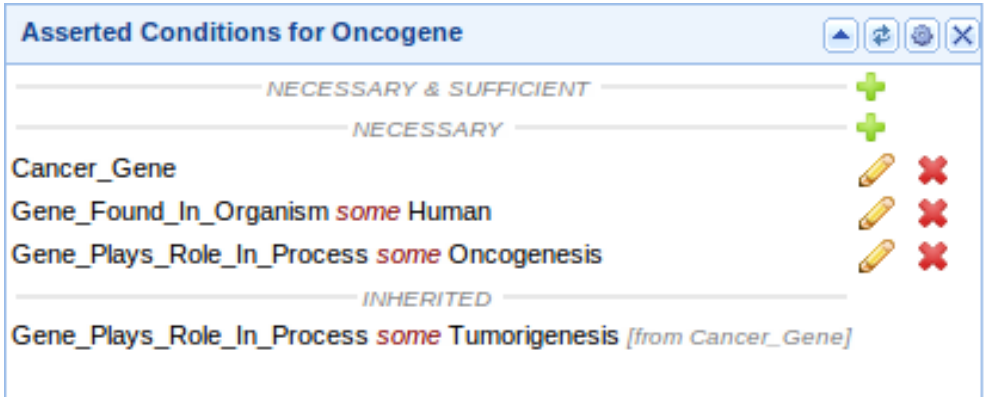
3 Large Scale Biomedical Projects

- The National Cancer Institute (NCI) Thesaurus
- The 11th Revision of the International Classification of Diseases (ICD-11)
- The Biomedical Resource Ontology (BRO)

The NCI Thesaurus

- Developed by the National Cancer Institute (NCI)
- Reference vocabulary for clinical care, translational and basic research and cancer biology
- Over 80.000 classes
- Rich OWL representation



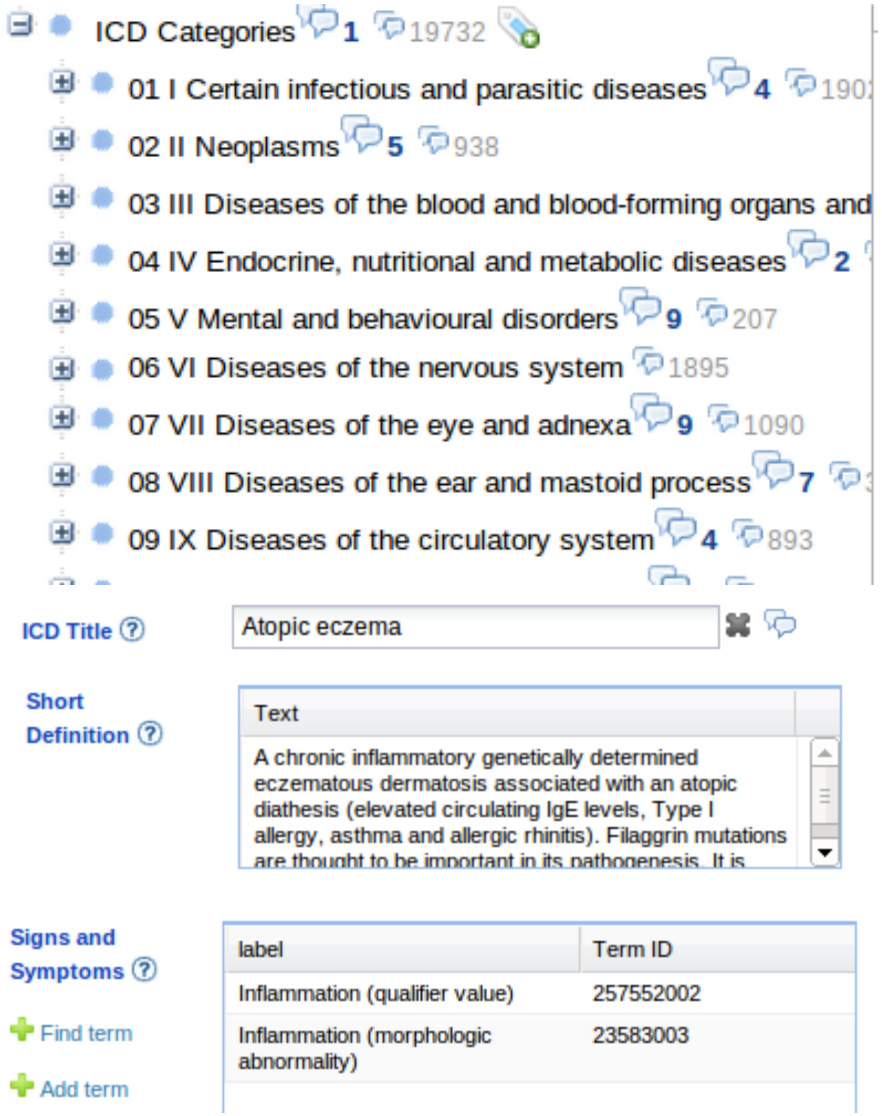


The NCI Thesaurus (cont.)

- Well defined workflow enforced in the tool
- 20 editors
- Lead editor accepts or rejects changes
- New versions published regularly (monthly)

The International Classification of Diseases (ICD)

- Standard diagnostic classification for epidemiology, health management and clinical use
- Developed by the World Health Organization (WHO)
- Used in all United Nations Countries for health statistics, to monitor health care spending, policy making
- 11th revision is on going
- Over 20.000 diseases

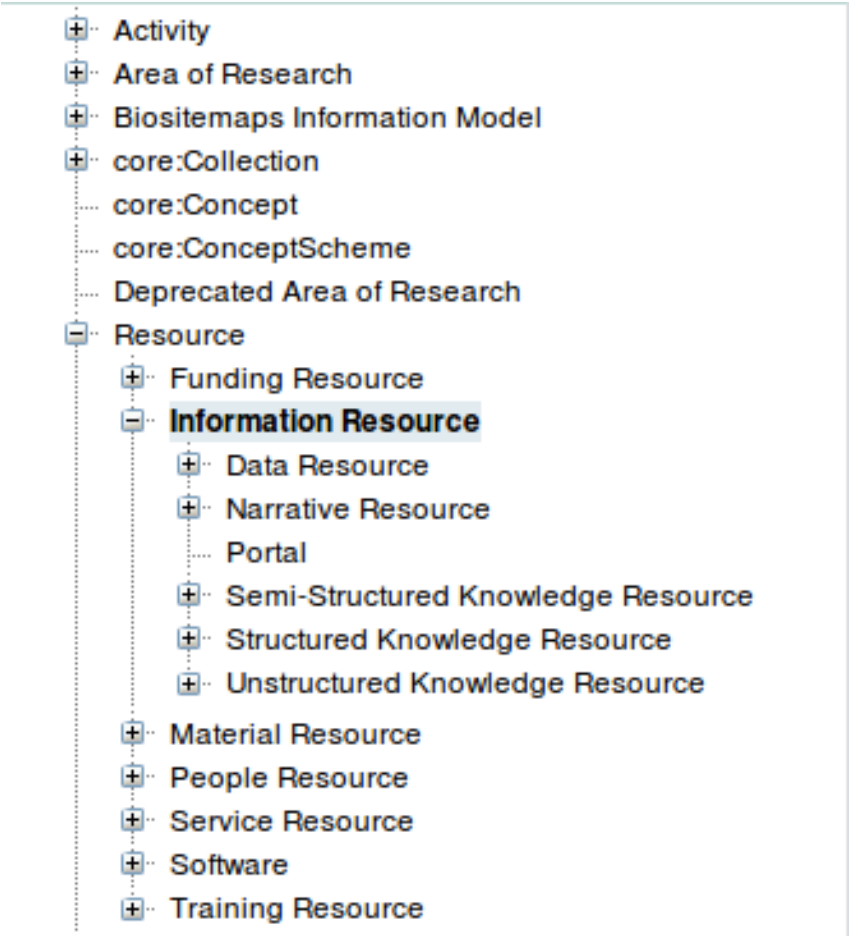


ICD-11

- Represented in OWL
- Core ontology developed by the Health Informatics Modeling Group (HIM-TAG)
- Content filled in by international domain experts in a web-based platform (iCAT – customization of WebProtégé)
- Workflow still being defined (meanwhile KA goes on)
- Early stages of development

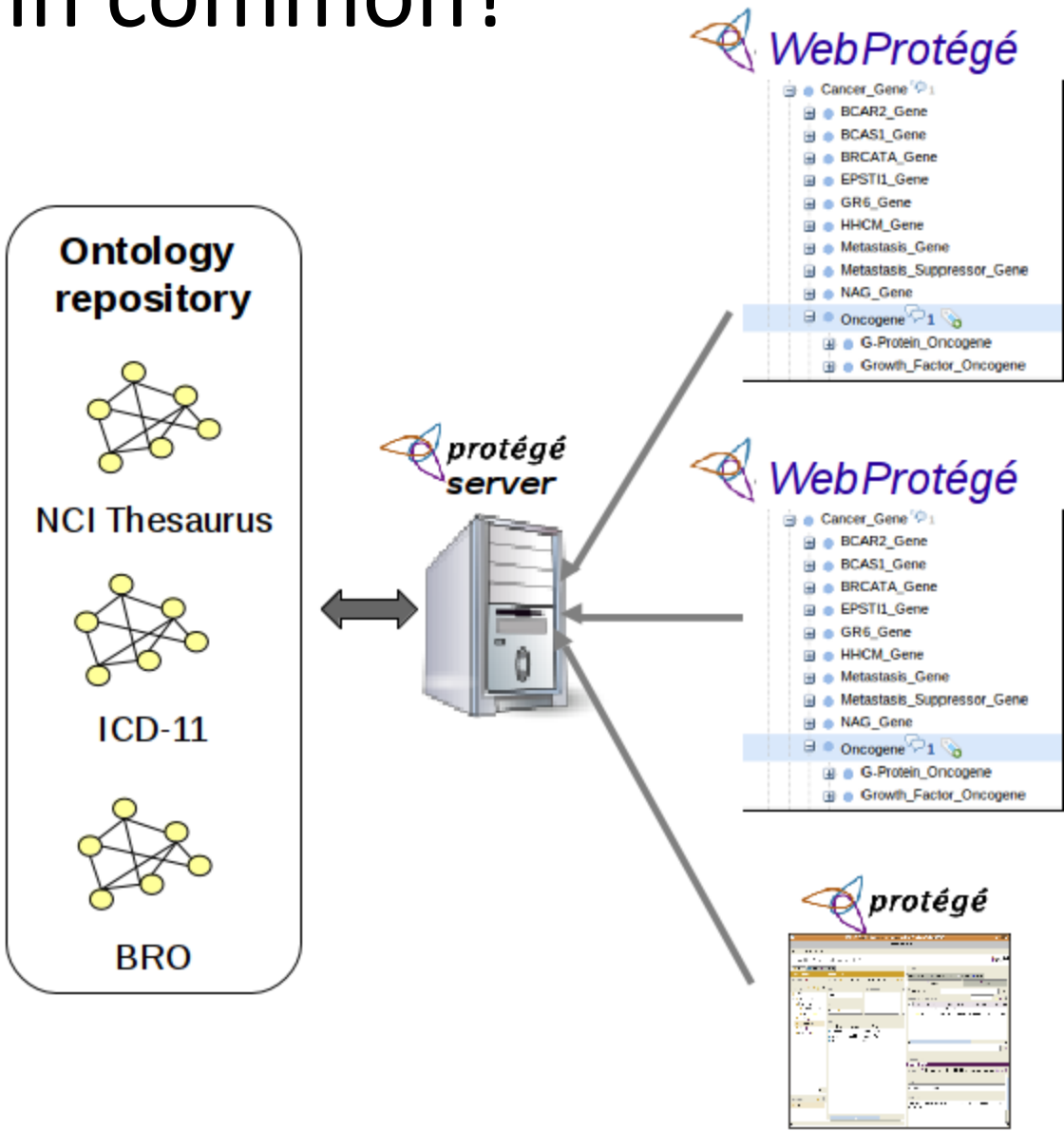
The Biomedical Resource Ontology (BRO)

- Biositemap project
- NIH National Center for Biomedical Computing
- Enable researchers in biomedicine to publish metadata about biomedical data, tools and services
- Controlled vocabulary
- RDF(S)
- Small group of editors
- Web-based tool to edit and carry out discussions

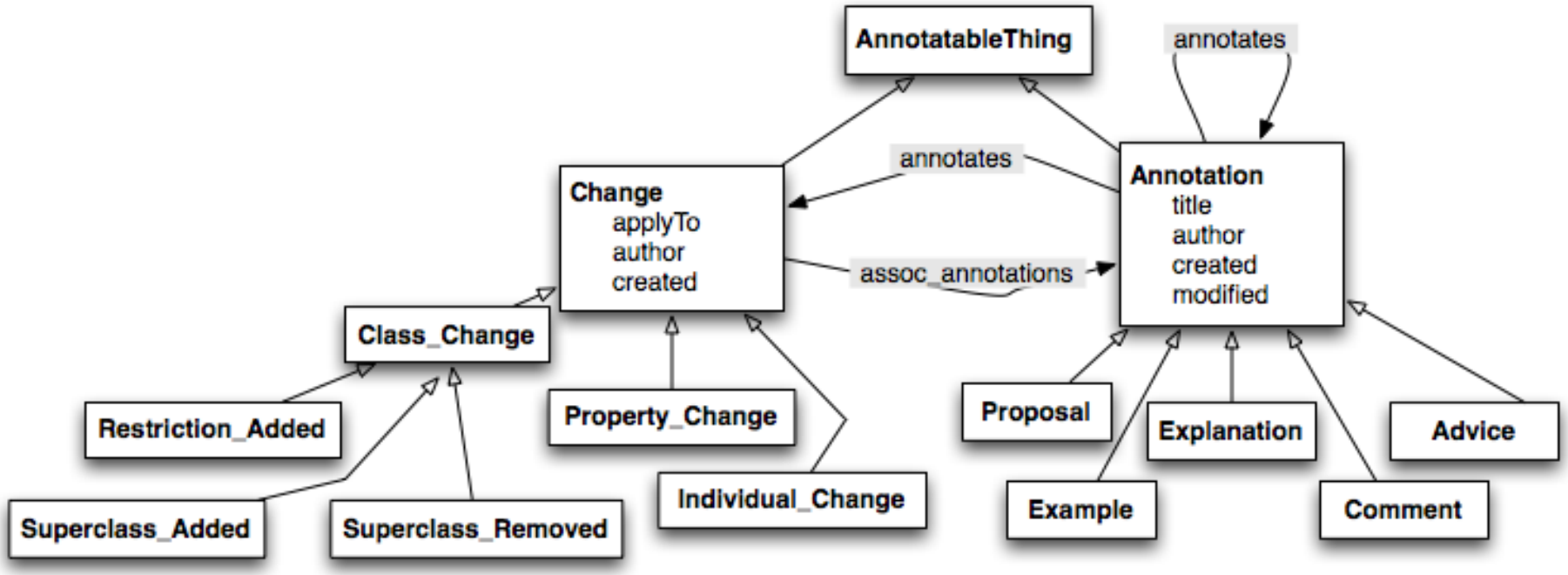


What do these projects have in common?

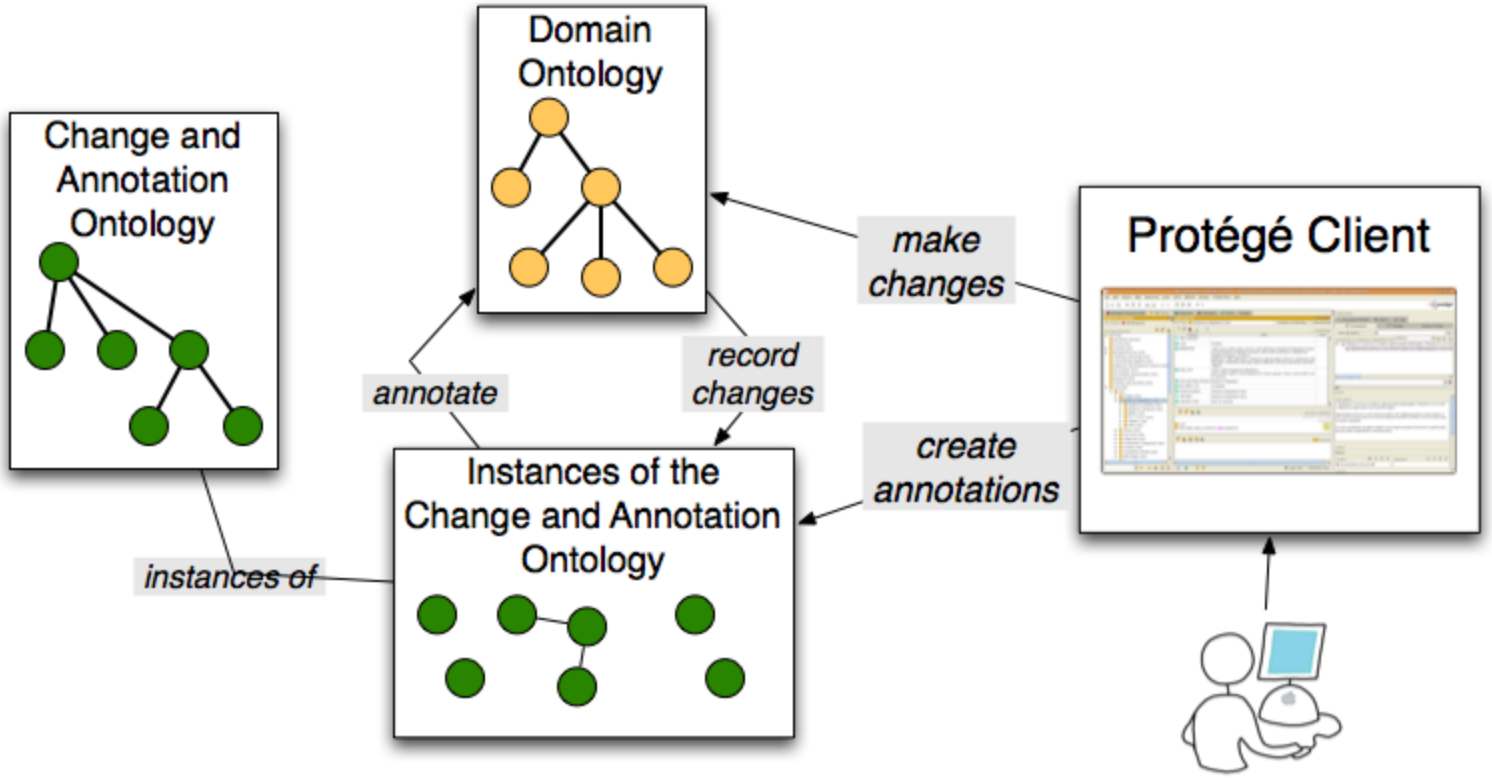
- Developed with Collaborative Protégé
- Client-server model with support for:
- **Change tracking**
- **Notes and discussions**
- Changes and notes stored as instances in the Changes and Annotation Ontology (ChAO)



The Changes and Annotation Ontology (ChAO)



Storing Changes and Notes in ChAO



What we used for our analysis

- Two types of collaboration activity
- **Change logs**: who, what and when did it
- **Notes**: why they did it, and other discussions
- Recorded as instances in ChAO

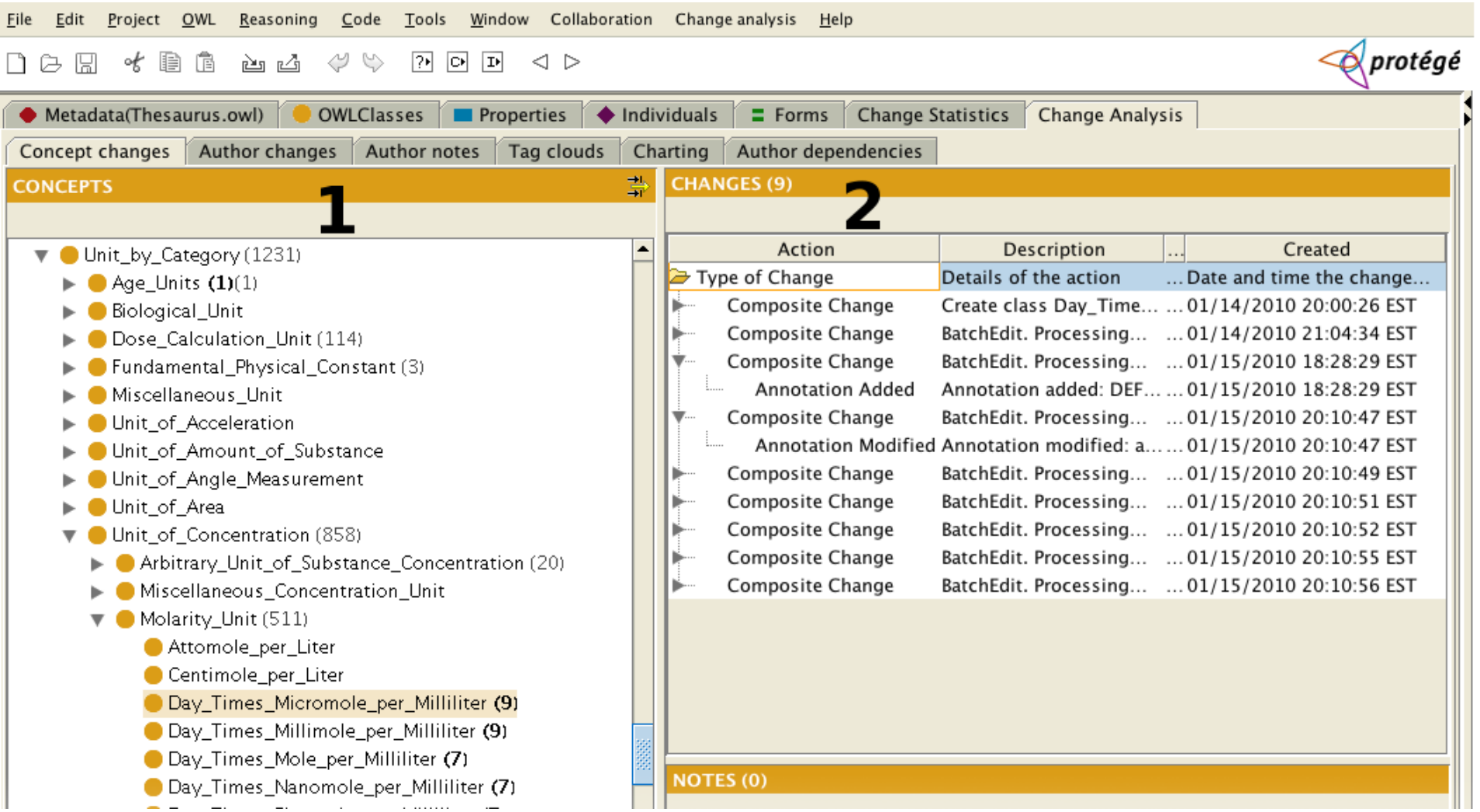
The collaboration data

Ontology	Dates	No. of changes	No. of notes	No. of authors
NCI Thesaurus	October 2009 – April 2010	43.702	0	10
ICD-11	November 2009 – May 2010	14.554	4.768	19
BRO	February 2010 – March 2010	762	373	5

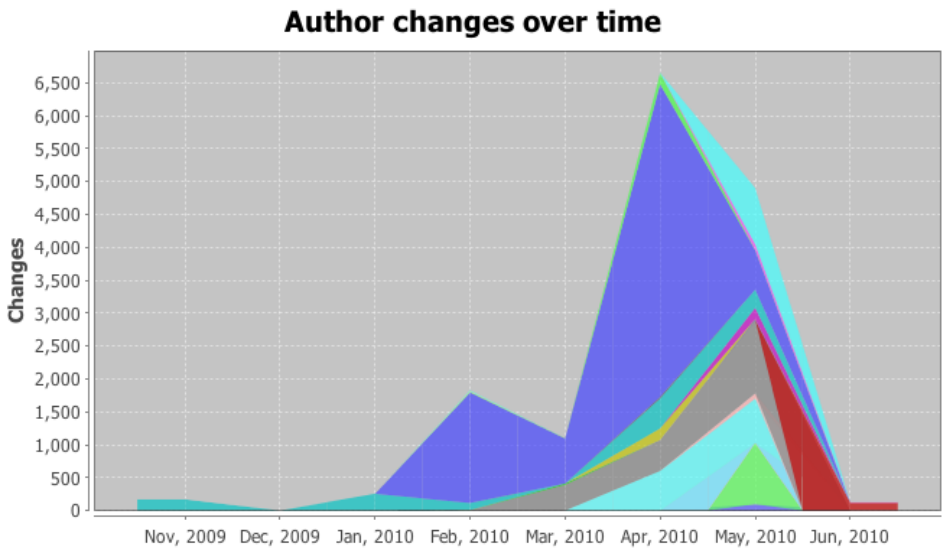
The Change Analysis Plugin

- Protégé plugin for **management of collaborative ontology development projects**
- Shows **statistics and aggregations of changes and notes** from ChAO:
  - Concept changes view
  - Changes over time per author
  - Author dependency network
  - Tag clouds for changes and notes
  - Similar notes views
- Used the plugin to perform the **qualitative analysis** of the data

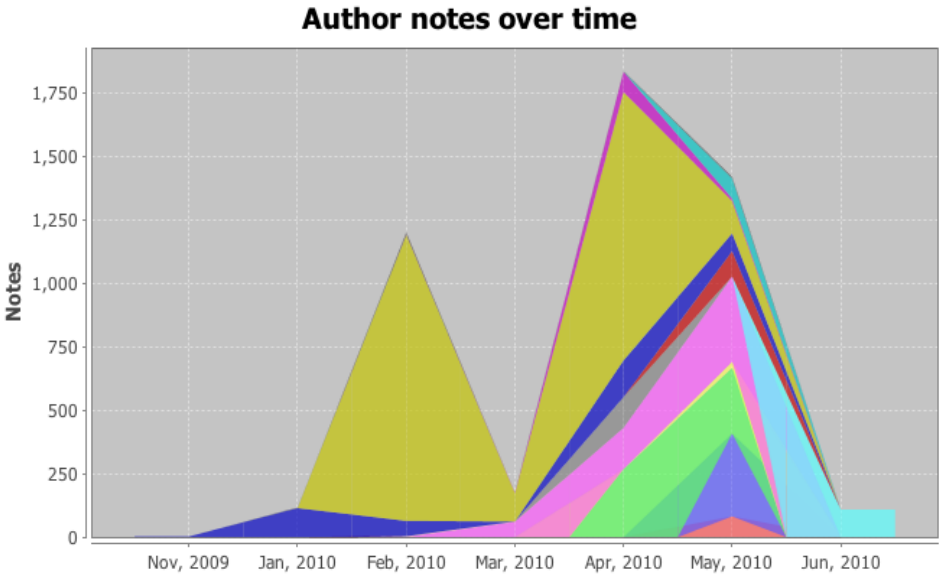
Concept changes view



Change and Note Statistics for ICD-11 (June 2010)

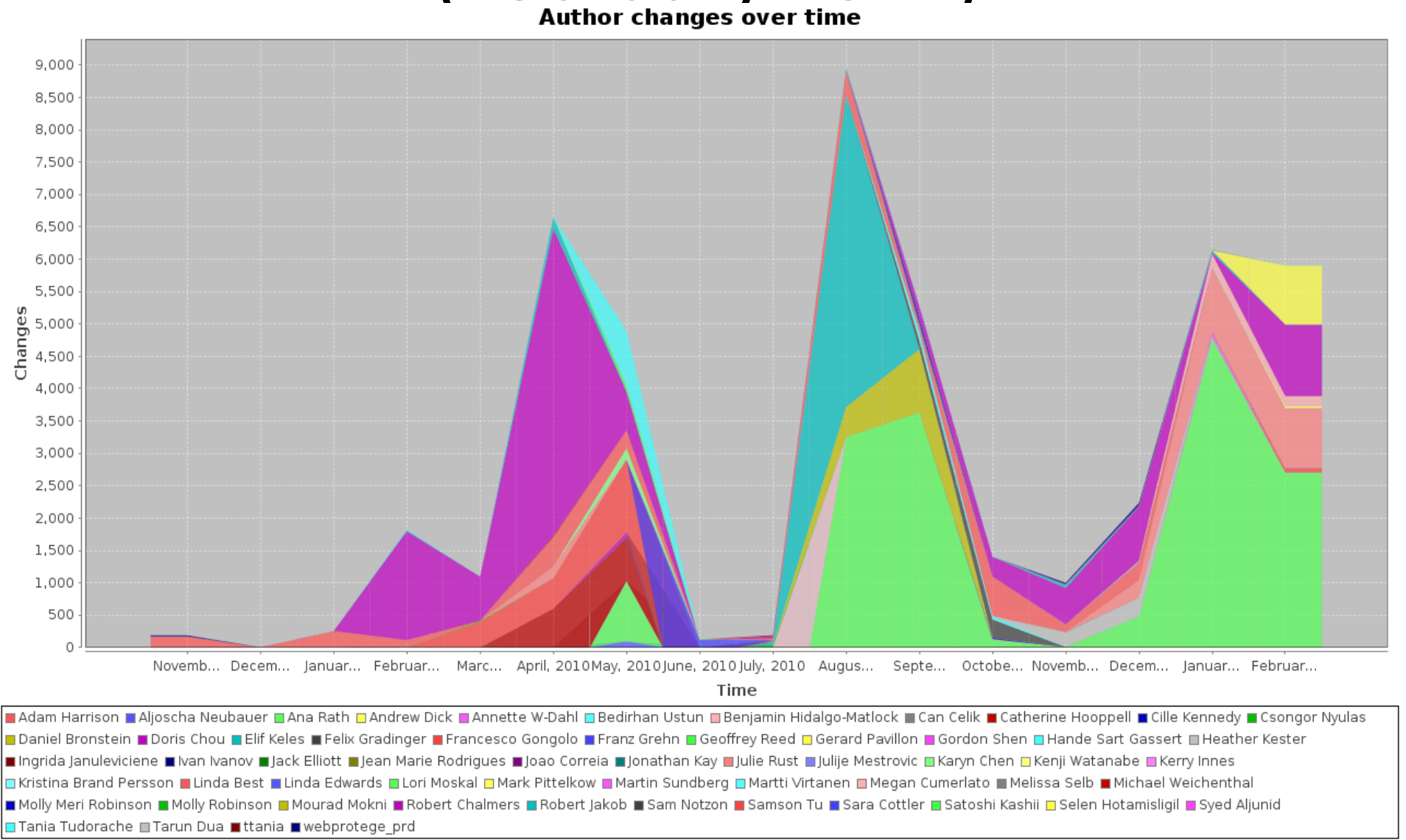


(a) Change contributions.

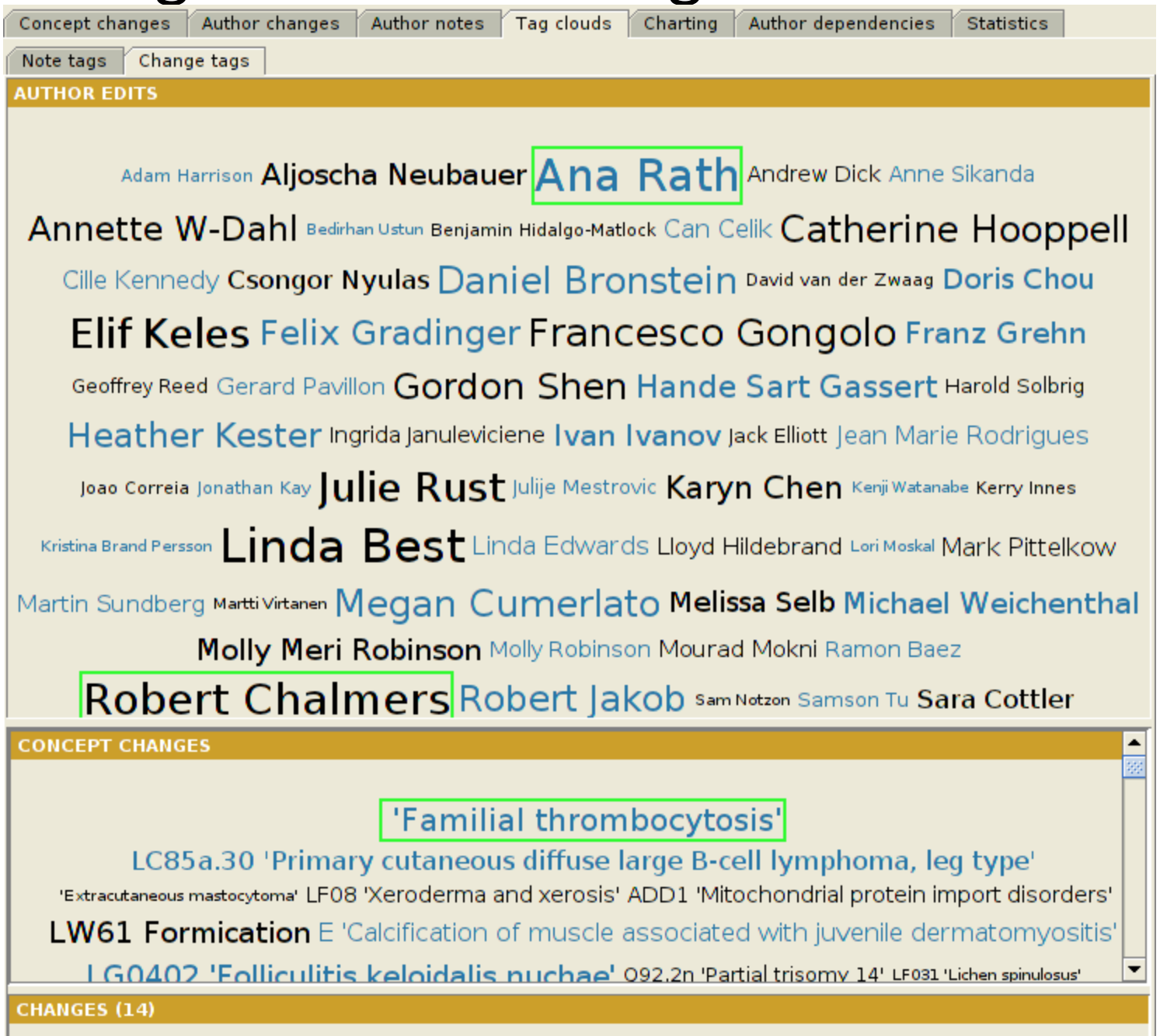


(b) Note contributions.

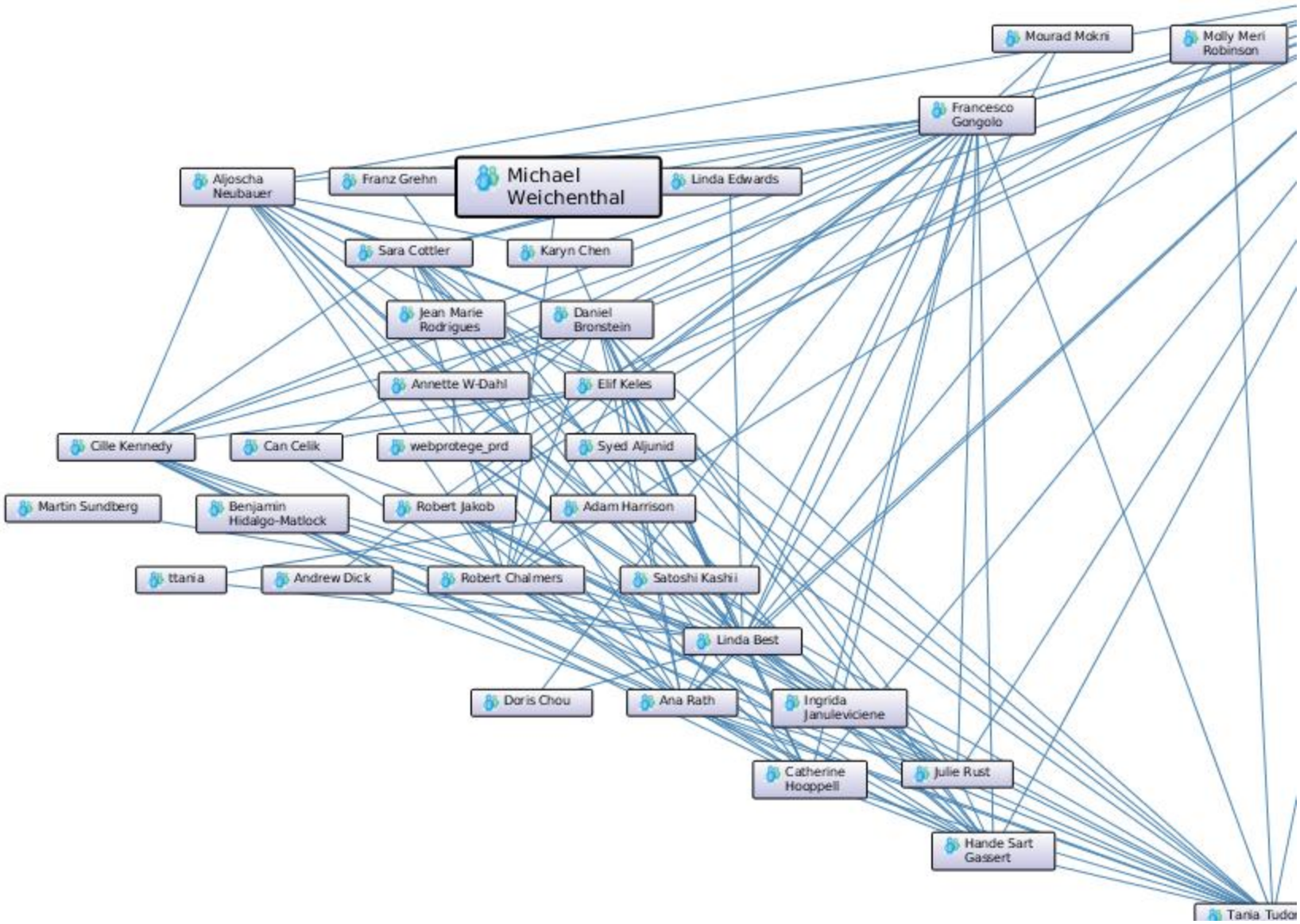
Changes statistics ICD-11 (February 2011)



Tag clouds for changes in ICD-11



Author dependency network ICD-11



Author dependencies

- Nodes in graph are authors
- Edges represent implicit or explicit dependencies
- **Explicit dependency**: two authors have edited the same entity
- **Implicit dependency**: two authors have edited entities related by a subclass or object property relation (or n-links away)
- Graph: Social network describing the relationship between the authors and potentially overlapping changes

Qualitative findings

- Try to extract patterns of consistent author behavior

1. We could differentiate the authors by **where in the class hierarchy** they were making changes
2. We analyzed the **overlap between the authors and changes** they made
3. We could differentiated authors by **types of changes** they made

- Try to associate the patterns of behavior with certain author roles in the project

Identifying user roles – Methodology

- Author role = **set of expected behaviors**

1. Created a **feature-vector representation** of an author based on previous observations
2. Applied **clustering to derive logical groupings** for the authors
3. Applied **statistical analysis** to determine what **characteristics** make each cluster **unique**

1. Feature-vector representation of authors

$$\vec{a} = (C_{del}, C_{add}, C_{mov}, C_{pro}, M, L, D, O, CE)$$

Table 1: Summary of author features used in author vector representation.

Symbol	*Feature*	*Explanation*
C_{del}	Deletion	Ratio of deletion changes to all changes committed.
C_{add}	Addition	Ratio of terms added to all changes committed.
C_{mov}	Move	Ratio of terms moved to all changes committed.
C_{pro}	Property change	Ratio of terms where a property has been added/modified to all changes committed.
M	Multi-author change	Number of times an author edits a term that is also edited by another author divided by the total number of terms modified by the author.
L	Leaf changes	Number of times an author edits a leaf concept term divided by the total number of terms modified by the author.
D	Relative depth	The average depth of a term change relative to the average depth of any given concept term in the ontology.
O	One hierarchy	The percentage of changes an author makes that are restricted to one level of the ontology hierarchy.
CE	Centrality	The average centrality of an author.

Clustering (2) and statistical analysis (3)

2. Clustering:
   - Discarded authors with less than 10 changes (ICD-11)
   - Applied repeated K-means clustering to divide authors into different groups based on their similarity (k=2 -> 8)
   - Evaluated the quality of clusters
   - > k = 5 – optimal number of clusters

3. Statistical analysis
   - Determine the characteristics that made each cluster unique, we applied multiple Analysis of Variance (ANOVA)
   - Determine statistically relevant features of each cluster

Results

- **Statistically significant differences ($p < 0.05$) across all features** except the multi-author edit feature
- 0 – less activity, 1 – more activity
- Relative depth: > 1 – deeper in hierarchy, < 1 higher in hierarchy
- To pinpoint differences, used Tukey range test -> **determine set of features that characterized each cluster**
- Discovered **5 author roles** with distinct characteristics

2: Feature comparison of means across all five clusters. Features correspond to those described in Table 1.

	Means					
Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	***p*-value**
Deletion	0.068	0.436	0.0858	0.040	0.077	0.023
Addition	0.352	0.220	0.800	0.464	0.270	< 0.001
Move	0.256	0.010	0.018	0.189	0.005	< 0.001
Property change	0.307	0.242	0.238	0.265	0.757	0.001
Multi-author change	0.528	0.469	0.127	0.051	0.328	0.060
Leaf changes	0.443	0.818	0.783	0.628	0.687	0.001
Relative depth	0.833	1.35	1.58	1.33	0.591	0.001
Centrality	0.889	0.448	0.537	0.753	0.973	< 0.001
One hierarchy	0.470	0.970	0.936	0.855	0.543	< 0.001

Identified user roles

- Ontology expert
- Content manager
- Domain expert
- Central domain expert
- Content editor

Table 3: Summary of roles

Cluster	*Role*	*Characteristics*	*Primary activity*	*Size*
Cluster 1	Ontology expert	Highly central author, makes changes over multiple hierarchies, involved in fewer leaf changes than domain experts, but performs a lot of movement changes in the hierarchy.	Organizational	4 authors: 3 ICD 1 BRO
Cluster 2	Content manager	Edits mostly in one sub-hierarchy, low centrality, performs few movement changes, but a high number of deletions.	Hierarchy clean-up	4 authors: 3 ICD 1 NCI
Cluster 3	Domain expert	Edits mostly deep within one hierarchy, low centrality, few moves, but lots of concept additions.	Content creation	12 authors: 5 ICD 7 NCI
Cluster 4	Central domain expert	Edits are restricted primarily to one sub-hierarchy, however, unlike domain experts, these authors are much more central and their changes occur at a higher level in the hierarchy. They also perform more movement operations than domain experts.	Management and content creation of a specific area of the ontology	2 authors: 2 ICD
Cluster 5	Content editor	Highly central author, makes changes over multiple hierarchies, lots of leaf changes, and a high number of property changes.	Editing of existing content	6 authors: 2 NCI 4 BRO

Collaboration and changes

- Also analyzed the notes activity
- Q1: Is there a relationship between changes and discussions w.r.t a specific ontology term?
- Q2: Do people who make a lot of changes also participate in a lot of discussions?

Method and Results

- **Method**:
  - Quantitative analysis to measure correlation between changes and discussions
  - Used Pearson correlation coefficient (+1 – positive correlation; -1 – negative correlation)

  Q1:

    - **(i) Binary change and note vectors for each term**
    - **(ii) Count of changes and notes vector**

- **Results** (p-value < 0.001) :
  - **ICD-11**: (i) Binary: **0.841** (ii) Count: **0.543**
  - **BRO**: (i) Binary: **0.274** (ii) Count: **0.258**

Method and Results (cont.)

- Q2: Do people who make a lot of changes also participate in a lot of discussions?
- We compared the correlation between no. of changes an author makes and the no. of notes he creates
- **ICD-11: high correlation** – 0.953 (p-value < 0.001)
- **BRO: no correlation**

Discussion

- Domain experts edit mainly within a single hierarchy
- Ontology experts, content editors: all over the place
- Implications for tool builders and workflow design, e.g.:
- Support a **role-based editing workflow** that displays only relevant parts of the ontology to the domain expert

Discussion (cont.)

- Measure the **degree of interest** (DOI) for an author based on the historical change data
- Predictive measure about topics a user is interested in
- Use change data to **filter out** or **highlight important** parts of the ontology
- Reduce ontology load times and memory consumption
- Distinguish different topic areas of the ontology -> use in **modularization** of the ontology

Discussion (cont.)

- **Similarities** between collaborative ontology development and the open source software (OSS) and Wikipedia communities
- **OSS**:
  - *Project leaders, Core developers, Co-developers, and Active Users*
  - Strong relationship between mailing list activity and development activity
- **Wikipedia** (we used the same clustering method):
  - *All round editors, Watchdogs, Starters, Content Justifiers, Copy Editors, and Cleaners*

Conclusions and future work

- **First analytical study** of user roles and connection between change and discussion activity for collaborative ontology development
- Results: **clearly discernible roles**, **recommendations** for ontology tool builders
- Current limitation: small set of projects, limited set of authors
- Future: more data from current and other projects, and other tools
- Investigate relationship between **changes, collaboration and contribution quality**

An Analysis of Collaborative Patterns in Large-Scale Ontology Development Projects

Sean Falconer, Tania Tudorache, Natasha Noy
Stanford Center for Biomedical Informatics Research

K-CAP 2011
Banff, Alberta, Canada

June 27, 2011